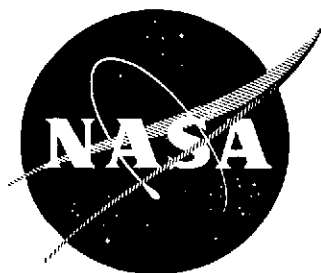


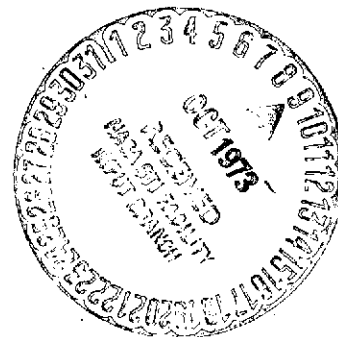
NASA CR- 134029
ERIM 31650-148-T



ESTIMATING CROP ACREAGE FROM SPACE-SIMULATED MULTISPECTRAL SCANNER DATA

by

R. F. Nalepka and P. D. Hyde
INFRARED AND OPTICS DIVISION



prepared for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Lyndon B. Johnson Space Center
NAS 9-9784 Task B2.9
Earth Observations Division

(NASA-CR-134029) ESTIMATING CROP ACREAGE
FROM SPACE-SIMULATED MULTISPECTRAL SCANNER
DATA (Environmental Research Inst. of
Michigan) —53 p HC \$4.75 CSCL 02C

N73-31359

Unclass

G3/13 14651

52

NOTICES

Sponsorship. The work reported herein was conducted by the Environmental Research Institute of Michigan (formerly the Willow Run Laboratories of The University of Michigan's Institute of Science and Technology) for the National Aeronautics and Space Administration, Johnson Space Center, under contract NAS 9-9784, Task B2.9; the Technical Monitor is Dr. Andrew Potter (TF3).

Disclaimers. This report was prepared as an account of Government-sponsored work. Neither the United States, nor the National Aeronautics and Space Administration (NASA), nor any person acting on behalf of NASA:

- (A) Makes any warranty or representation, expressed or implied with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or
- (B) Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used above, "person acting on behalf of NASA" includes any employee or contractor of NASA, or employee of such contractor, to the extent that such employee or contractor of NASA or employee of such contractor prepares, disseminates, or provides access to any information pursuant to his employment or contract with NASA, or his employment with such contractor.

Availability Notice. Requests for copies of this report should be referred to:

National Aeronautics and Space Administration
Scientific and Technical Information Facility
P. O. Box 33
College Park, Maryland 20740

Final Disposition. After this document has served its purpose, it may be destroyed. Please do not return it to the Environmental Research Institute of Michigan.

Technical Report

ESTIMATING CROP ACREAGE FROM SPACE-SIMULATED MULTISPECTRAL SCANNER DATA

by

R. F. Nalepka and P. D. Hyde
INFRARED AND OPTICS DIVISION



prepared for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

August 1973

NAS 9-9784

Lyndon B. Johnson Space Center
Houston, Texas
Earth Observations Division



FOREWORD

This report describes part of a comprehensive and continuing program of research in multispectral remote sensing of environment from aircraft and satellites. The research is being carried out for the NASA Lyndon B. Johnson Space Center, Houston, Texas, by the Environmental Research Institute of Michigan (formerly the Willow Run Laboratories, a unit of The University of Michigan's Institute of Science and Technology). The basic objective of this program is to develop remote sensing as a practical tool for obtaining extensive environmental information quickly and economically.

In recent times, many new applications of multispectral sensing have come into being. These include agricultural census-taking, detection of diseased plants, urban land studies, measurement of water depth, studies of air and water pollution, and general assessment of land-use patterns. Yet the techniques employed remain limited by the resolution capability of a multispectral scanner. Techniques described in this report may help to overcome this limitation by enabling either examination of the contents of a given scanner resolution cell or faster estimation of the contents of a larger area.

To date, our work on estimation of proportions has included: (1) extension of the signature concept to a mixture of objects; (2) development of a statistical and geometric model for sets and mixtures of signatures; (3) evaluation of computational methods used to estimate proportions of a mixture by maximum likelihood; (4) measurement of computer time required by various algorithms; (5) creation of a computational technique for assessing the expected accuracy of estimation as a function of the signature set; (6) development of techniques to identify alien objects; (7) testing and evaluating the proportion estimation algorithms on artificial as well as actual multispectral scanner data; and (8) examining the problem of establishing signatures when pure samples of the objects of interest are not available.

The research covered in this report was performed under Contract NAS9-9784, Task B2.9, and covers the period from November 1971 through January 1973. Dr. Andrew Potter has been Technical Monitor. The program was directed by R. R. Legault, Associate Director of the Environmental Research Institute of Michigan (ERIM), and by J. D. Erickson, Principal Investigator and Head of the ERIM Multispectral Analysis Section. The ERIM number for this report is 31650-148-T.

ACKNOWLEDGMENTS

The authors wish to acknowledge the direction provided by Mr. R. R. Legault and Dr. Jon D. Erickson. Significant technical contributions were made by Dr. H. M. Horwitz and helpful comments were supplied by Dr. R. B. Crane. The authors also thank Mrs. L. A. Parker for her efforts in preparing this report and for secretarial assistance since contract inception.

ABSTRACT

The need for multispectral data processing methods to permit the estimation of proportions of objects and materials appearing within the instantaneous field of view of a scanning system is discussed. An algorithm developed for proportion estimation is described as well as other supporting processing techniques. Application of this algorithm to space-simulated multispectral scanner data is discussed and some results presented and compared. Results reported herein indicate that, for this data set, the true proportions of the various crops contained within this data set are with one exception more closely in agreement with the proportions determined by the proportion estimation algorithm than with the proportions determined by conventional classification algorithm.

CONTENTS

Foreword	iii
Acknowledgments	iv
Abstract	v
List of Figures	viii
List of Tables	viii
1. Summary	1
2. Introduction	3
3. Development of the Proportion Estimation Capability	11
3.1. Model for Signatures of Mixtures	11
3.2. Estimation of Proportions	12
3.2.1. Solution by Quadratic Programming	13
3.3. Detection of Alien Objects	14
3.3.1. Hyperplane Criterion	15
3.3.2. Out-of-Plane Test	15
3.3.3. Alien Object Test Summary	17
3.4. Description of the Proportion Estimation Program	17
3.5. Signature Analysis	18
4. Application of the Proportion Estimation Capability	23
4.1. Data Preparation	23
4.2. Signature Extraction	25
4.3. Data Processing and Evaluation	26
5. Extracting Signatures from Low Resolution Data	33
6. Conclusions and Recommendations	35
Appendix I: Number of Samples Required for Training	37
Appendix II: Details of the Alien Object Tests	40
References	45
Distribution List	46

FIGURES

1. Illustration of the Mixtures Problem at Field Boundaries	4
2. Reflectance Spectra of Individual Objects	5
3. Reflectance Spectra of Mixtures of Objects	6
4. Effect of Field Area and Shape on Multispectral Training and Classification for 300×300 ft Spatial Resolution	8
5. The Extended Signature Simplex	16
6. Examples of Standard and Estimated Proportion Classification Maps . . .	19
7. Signature Simplex with Unit Contour Ellipsoids	21
8. Ill-Conditioned Signature Simplex	21
9. Composition of a Portion (Area 1) of Space-Simulated Lenawee County Data as Determined by Four Methods	30
10. Composition of a Portion (Area 2) of Space-Simulated Lenawee County Data as Determined by Four Methods	31
I-1. Inaccuracy in the Estimates of Signatures as a Function of the Number of Samples	39

TABLES

1. Spectral Sensitivity of the Michigan Multispectral Scanner for August 1970	24
2. Results of Estimating Proportions on Training Areas of 10K ft, August 1970 Lenawee County Data	27
3. Proportions of Materials for August 1970 Lenawee County Data Set Using Three Classification Approaches after Combining Data Points to Simulate 300×300 ft Resolution	29

ESTIMATING CROP ACREAGE FROM SPACE-SIMULATED MULTISPECTRAL SCANNER DATA

1

SUMMARY

There are many potential applications for multispectral remote sensing. Under certain circumstances, however, the applications are limited by the spatial resolution of the sensing device. To help overcome this limitation, techniques have been developed for estimating the proportions of the objects contained within each pixel or resolution element of a multispectral scanner. Such techniques were tested on artificial data in 1971 [1].

The investigations reported herein had three objectives: (1) to develop additional supporting techniques for application of the ERIM proportion estimation algorithm, (2) to test this algorithm on actual multispectral scanner data, and (3) to demonstrate the superiority of proportion estimation over standard classification for estimating crop acreage or area from space-altitude multispectral scanner data.

As a part of objective (1), a statistical test was developed to detect resolution elements containing alien or unknown material. This test, based on the geometry of the signature simplex, helps eliminate from consideration data which would otherwise corrupt estimates of proportions.

Some work was done on estimating signature means and covariances from data in which no pure samples are to be had for training purposes. This capability is important because of the difficulty in finding pure training sets in low resolution data. A mathematical procedure was proposed for such estimation. However, in view of possible data handling difficulties, this problem merits further investigation.

To satisfy objectives (2) and (3) above, multispectral scanner data were obtained from a flight made at 10,000 ft over an agricultural area in Lenawee County, Michigan. The data were digitized and various signatures generated from a number of training areas that provided pure data samples of six major crops. Various subsets of these signatures were evaluated by using a geometric signature analysis technique; then for each crop, a subset of signatures was combined to yield a final crop signature. Simulated space data were generated by smoothing (i.e., data point averaging) the 10,000 ft data over the scan line and along the flight line to simulate the 300 ft by 300 ft spatial resolution expected from the ERTS and SKYLAB multispectral scanners. The signatures generated from the unsmoothed data were used to estimate proportions in the space-simulated data via the method of Theil and van de Panne [2]. Proportion maps

and an alien object map were produced for the entire Lenawee County flight line. Classification maps were also generated using the conventional, one-class-per-pixel approach.

These results were analyzed and compared with the ground truth information available for this flight line. Analysis indicated that, for those test areas where adequate and reliable ground truth is available, acreages determined by proportion estimation were closer to the truth than those determined by conventional classification procedures.

2

INTRODUCTION

In recent years, remote multispectral data collection and automatic processing techniques have proven feasible for many user applications in the fields of earth resource survey and management. For some applications, however, the spatial resolution limits of multispectral scanners restrict the usefulness of data so obtained.

Examples of this constraint are seen in agricultural resource management where it is important to know the number of acres planted to each of the major crops. Spaceborne multispectral scanning systems and automatic data processing schemes have been recommended as efficient means for providing the necessary information. However, the information that can be extracted via conventional processing techniques may not be sufficiently accurate, since the instantaneous ground patch viewed from space altitudes by the scanner comprises a significant fraction of typical agricultural fields. As a result many pixels overlap the common boundaries between adjoining fields.

To roughly determine the number of pixels which may reside on a field boundary, let us consider the idealized scene of Fig. 1. Here we see nine fields, as bounded by the solid lines, and a superimposed matrix of squares formed by the dashed lines with each such square defining the ground area of a single pixel. The illustration includes only enough pixels to cover the center field in the scene. Clearly, many pixel areas overlap the boundaries of the center field. For the case illustrated, 12 of the 16 pixels cover areas lying both inside and outside the center field; of the total center field area, some 55% is included in these 12 pixels.

When more than one ground cover is viewed in any given pixel, the apparent reflectance spectrum is modified. Figures 2 and 3 illustrate this. In Fig. 2 the reflectance spectra are depicted as they would appear individually for corn and bare soil. But if the sensor were to view corn and bare soil simultaneously, the effective reflectance spectrum would be quite different. This is shown in Fig. 3 for the combinations 20% corn/80% bare soil and 50% corn/50% bare soil where the spectra are simply weighted combinations of the pure spectra illustrated in Fig. 2.

Hence it is clear that the spectra generated by a combination or mixture of two or more objects will differ, sometimes considerably, from the pure spectra. The resulting spectra will not be characteristic of either object class. Thus those pixels in Fig. 1 which overlap the field boundaries would most likely be misclassified; and as a consequence the area covered by the crop of the center field would be underestimated by 55%, assuming that all four pixels wholly within the field were correctly classified.

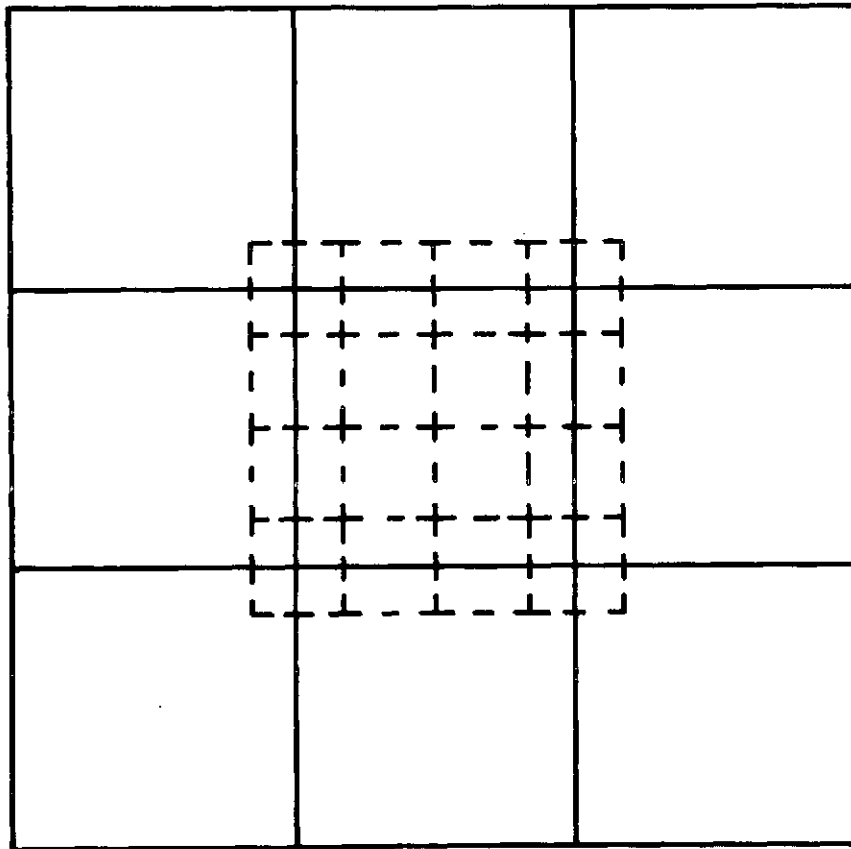


FIGURE 1. ILLUSTRATION OF THE MIXTURES PROBLEM AT FIELD BOUNDARIES

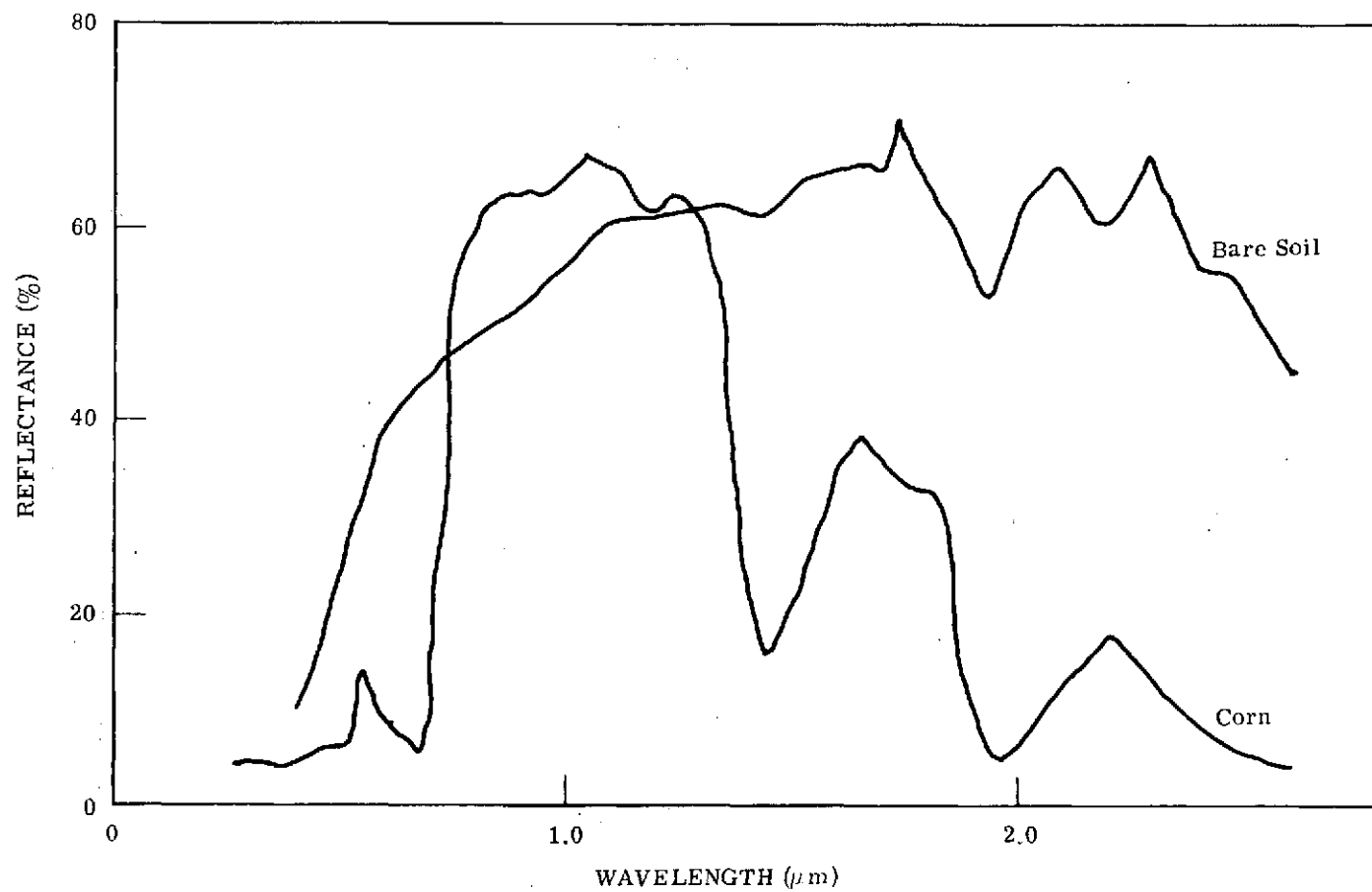


FIGURE 2. REFLECTANCE SPECTRA OF INDIVIDUAL OBJECTS

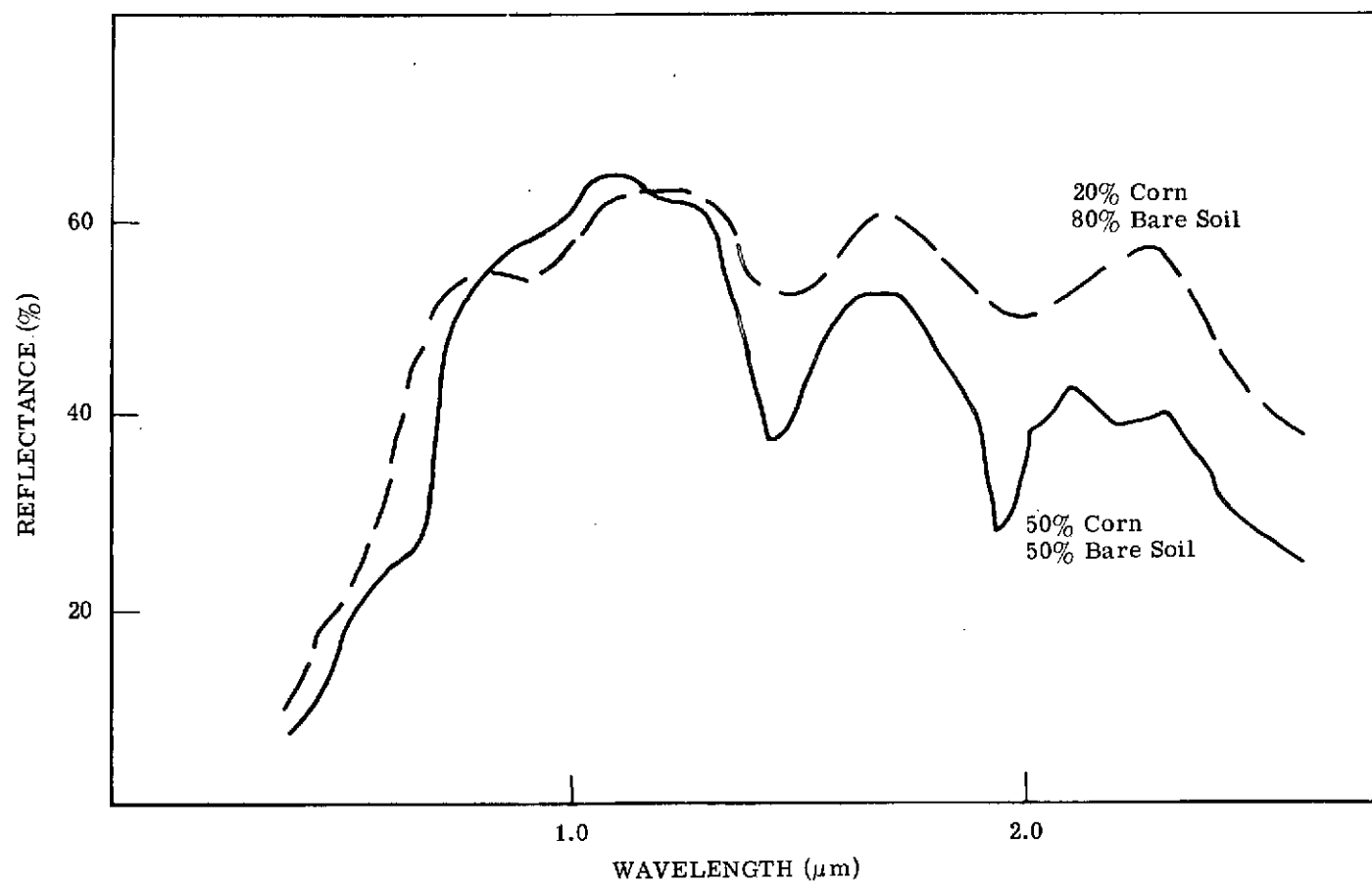


FIGURE 3. REFLECTANCE SPECTRA OF MIXTURES OF OBJECTS

We have carried out some calculations to roughly determine how serious this problem may be for spaceborne sensors viewing an instantaneous ground patch 300 ft square. Figure 4, which plots the results of these calculations, illustrates the effect of field area and shape on both the multispectral training and classification operations. This time we include rectangular fields as well as square ones. Dimensions of both the square and rectangular fields are integral multiples of 300 ft, and those pixels which overlap field boundaries are assumed to fall one-half or, at field corners, three-quarters outside the field. For the rectangular fields the small dimension is 600 ft. Square and rectangular fields of this sort represent the limiting conditions (best and worst) for regularly shaped areas and for the pixel arrangement described.

Two pairs of curves are plotted in Fig. 4. Those drawn as dashed lines relate the number of acres in a field to the number of 300 by 300 ft elements wholly within the field. This relationship becomes significant in training a recognition computer when the characteristic signatures of objects to be identified are being established. For each signature some minimum number of data points is required for adequate determination of signature statistics. (This is discussed in more detail in Appendix I.) As an example of the use of Fig. 4, assume that 40 samples are required. Using the dashed lines, we find that for 40 samples to be entirely within the boundaries of a single field, a square field of at least 110 acres would be required and a $2 \times n$ rectangular field of at least 170 acres. These field sizes are relatively large and fields of this size may not exist for all crops in a scene. This suggests the difficulties that are likely to arise in the training operation alone.

If training of the computer had already been accomplished, however, it would be desirable to know what errors might be expected when automatically determining the area covered by a selected set of object classes. Figure 4 shows these errors. The curves drawn as solid lines relate the number of 300 by 300 ft elements in the field to the percentage of the field area which is seen in combination with portions of adjoining fields. As previously mentioned, it is those pixels that overlap the field boundaries which would probably be classified incorrectly and thus produce the errors in computed acreage. For the 110-acre square field described above, 40 elements were totally within the field so a 25% error in the determination of the area of that field could be expected. For the 170-acre rectangular field a 51% error in field area could be expected. Because of the generally limited size farms in the Midwest, the fields are often rectangular in shape to permit more efficient planting and harvesting. Therefore the expected errors for the assumed pixel arrangement may well exceed the minimum value shown.

Not only are these errors of significant size, but in many regions of the country the field sizes considered above are somewhat larger than is typical. As shown in Fig. 4, even larger errors would result for smaller fields.

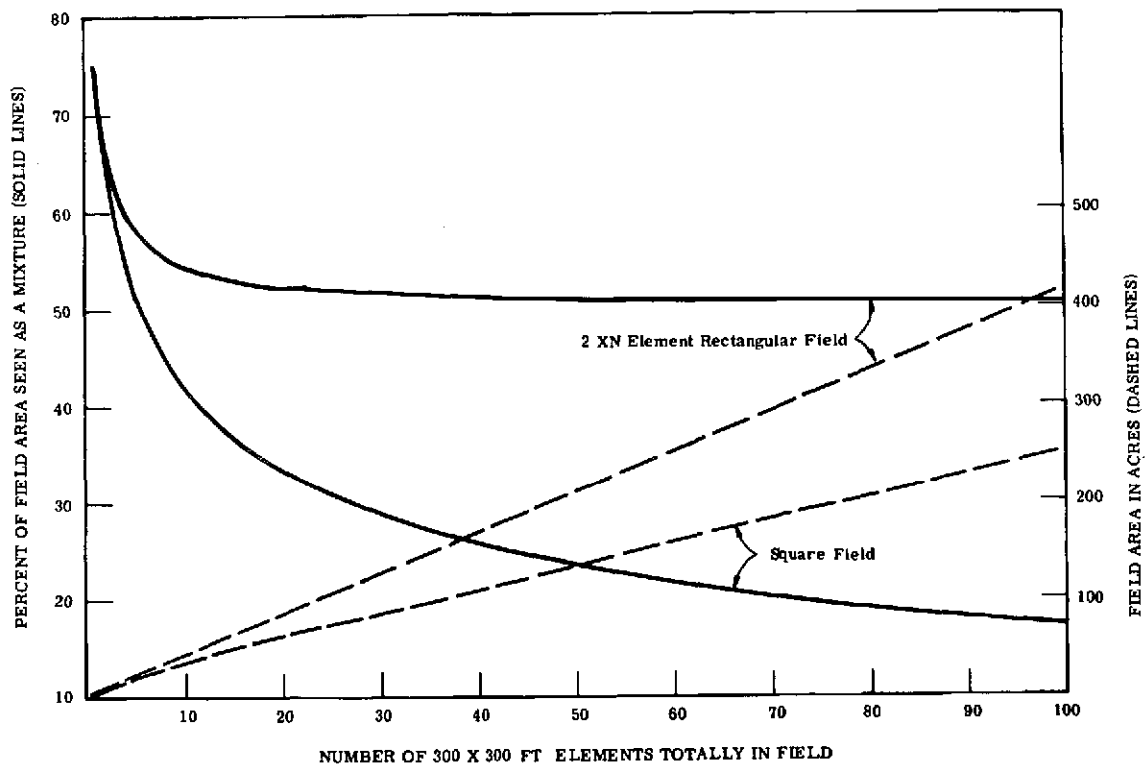


FIGURE 4. THE EFFECT OF FIELD AREA AND SHAPE ON MULTISPECTRAL TRAINING AND CLASSIFICATION FOR 300 x 300 FT SPATIAL RESOLUTION

If, as we have illustrated here, the percentage of pixels containing class boundaries is large, then a realistic estimation of acreage from low resolution scanner data may require a determination of the proportions of classes contained within the boundary pixels. The errors due to overlap might be reduced by increasing the system resolution, but this is not always feasible. It would seem easier to avoid this problem by estimating the proportions of the several materials in each pixel.

Toward developing suitable algorithms for estimating proportions in a mixture, considerable previous work has already been done [1, 3]. A simple Gaussian model was constructed for relating the multispectral signature of a mixture to the signatures of component materials. With this model, several existing computational methods were adapted for estimating the proportions by maximum likelihood. Special emphasis was placed on a method that assumes equal covariance matrices, since this simplifies the optimization problem and since this assumption can be justified practically and theoretically (Ref. 3, Section 3.4). Artificial scanner data were generated to test the computational methods for accuracy and speed. We found it possible to estimate proportions over an entire rectangular area directly from the average of data vectors from that area, with little or no decrease in accuracy and considerable increase in speed. Proportions estimated with and without averaging were compared. Also, a geometric criterion was developed to determine in advance whether good estimates could be obtained from a given set of signatures.

The present report, summarizing work accomplished to date, attempts to show the usefulness of the proportion-estimation method in a typical remote sensing application. Because the measurement of agricultural resources is a common application, an agricultural test area was chosen. And because we anticipated that estimation of proportions will be a necessary part of ERTS or SKYLAB measurement of earth resources, the estimation program was tested on simulated space data.

The next section (Section 3) describes the model used for mixture signatures and summarizes the computational methods required to solve the proportion-estimation problem. It further discusses the problem of detecting alien objects—that is, objects in the scene not represented by any of the given signatures. Statistical and geometrical criteria have been developed for detecting and eliminating data points that probably do not represent mixtures of the given materials (and hence could corrupt the estimated proportions).

Section 4 describes the preparation of a data set, the application of the proportion estimation to the data set, and an analysis of the results.

Section 5 deals with estimation of signature means and covariances from mixtures data. This type of estimation may be essential in obtaining signatures from space data in which very few data points may correspond to "pure" crops or surface compositions.

DEVELOPMENT OF THE PROPORTION ESTIMATION CAPABILITY

In this section our model for signatures is described and the mathematical procedures for estimation of proportions are defined. We describe the model and solution of the estimation problem by maximum likelihood. We then discuss the quadratic programming solution to the optimization problem derived from the maximum likelihood criterion, and also some of the algorithms considered. Finally we discuss the alien object problem and explain our method of geometric signature analysis.

3.1. MODEL FOR SIGNATURES OF MIXTURES

If the IFOV of a multispectral scanner is large compared to the structure of the scene being scanned, a single resolution cell or pixel may contain more than a single object or material. Suppose the scanner has n spectral channels and that the signature of object class i , where $1 \leq i \leq m$, is represented by an n -dimensional Gaussian distribution with mean A_i and covariance matrix M_i . If the proportion of object class i in the resolution cell is p_i , and p is the vector (p_1, \dots, p_m) , then let the mixture signature for this combination of classes have mean vector A_p and covariance matrix M_p .

To find expressions for A_p and M_p , consider the following model. If the resolution cell contains elements only of object class i , assume that it contains N_i elements. With each element, associate a random variable with mean A_i^* and covariance matrix M_i^* . We then have

$$A_i = N_i A_i^*$$

If we assume statistical independence for these N_i random variables, then we also have

$$M_i = N_i M_i^*$$

Now, if the proportion of the resolution cell covered by elements of object class i is p_i , then the number of elements of this type in the resolution cell is $p_i N_i$. Thus

$$A_p = \sum_i p_i N_i A_i^* = \sum_i p_i A_i$$

If we assume statistical independence for random variables associated with elements from different object classes, we also obtain

$$M_p = \sum_i p_i N_i M_i^* = \sum_i p_i M_i$$

Since the pure signatures of objects i are taken to be Gaussian distributions, the distribution associated with the proportion vector p is also Gaussian and may be defined by the mean and covariance,

$$A_p = \sum p_i A_i$$

$$M_p = \sum p_i M_i$$

These formulas constitute our model for signatures of object-class combinations (mixtures) in terms of signatures of the individual object classes.

3.2. ESTIMATION OF PROPORTIONS

The model for a mixture signature can now be used to estimate the proportion vector p for a mixture of materials in the resolution cell corresponding to a signal data vector from a multispectral scanner. This is done by minimizing the negative log of the likelihood function associated with the Gaussian distribution having mean A_p and covariance M_p . Let A_i be the signature mean vector for the i -th object class, and M_i the corresponding variance matrix. Let y be the n -dimensional data vector from the scanner. Then the proportion vector p is estimated by finding a value for p that minimizes

$$F(p) = \ln |M_p| + \left\langle y - A_p, M_p^{-1} (y - A_p) \right\rangle$$

subject to the constraint that

$$\sum p_i = 1 \text{ and } p_i \geq 0 \text{ for } 1 \leq i \leq m$$

where A_p and M_p are as defined above. Note that $|M|$ is the determinant of M , M^{-1} is its matrix inverse, and $\langle u, v \rangle$ denotes the inner product of vectors u and v . $F(p)$ is the negative natural log of the Gaussian density function with the constant term eliminated and with mean A_p and M_p evaluated at the point y .

In general, minimizing $F(p)$ subject to the given constraints is quite difficult, but it can be shown that the minimization problem is easier if the M_i are equal. Also, it can be shown that the optimal p is unique whenever the covariance matrices M_i are equal. Previous work with simulated data (Ref. 3) has shown that the equal covariance assumption leads to a good approximation to the true proportions, even in the cases where the M_i are not equal. For these reasons our previous work emphasized the case where the M_i are equal, and our present estimation algorithm also makes this assumption.

3.2.1. SOLUTION BY QUADRATIC PROGRAMMING

Suppose now that all covariances M_i are equal, so that $M_i = M$ for all i . Then the function $F(p)$ becomes

$$F(p) = \ln |M| + \left\langle y - A_p, M^{-1}(y - A_p) \right\rangle$$

The first term drops out of the optimization problem, since it is constant. If M is factored into the form

$$LL^T = M$$

then

$$\left\langle y - A_p, M^{-1}(y - A_p) \right\rangle = \left\langle y - A_p, L^{-T}L^{-1}(y - A_p) \right\rangle = \left\langle L^{-1}(y - A_p), L^{-1}(y - A_p) \right\rangle$$

Thus minimizing $F(p)$ is equivalent to minimizing

$$G(p) = \|Z - B_p\|^2$$

where $Z = L^{-1}y$ and $B_p = L^{-1}A_p$. This is geometrically equivalent to projecting B_p of Z onto the convex hull of the points $B_i = L^{-1}A_i$. With the constraints on p , the problem becomes one of quadratic programming and can be solved by any of several known methods.

In our previous work, three algorithms were considered for solving the quadratic programming problem: the Frank and Wolfe, Complementary Pivot, and Theil and van de Panne methods. A brief description is given below for the Theil and van de Panne method, which is the algorithm used in our present program. For descriptions of the other two methods, see Refs. 1 and 3.

The Theil and van de Panne method minimizes $G(p)$ by computing certain projections of Z onto hyperplanes determined by vertices B_i . (Usually one does not have to compute all such projections.) Let S be a subset of the index set $(1, \dots, m)$ and let π_S be the orthogonal projection of Z onto the hyperplane spanned by the B_i such that $i \in S$. Then there exists a proportion vector p^S which satisfies

$$\pi_S(Z) = \sum_i p_i^S B_i$$

$$\text{where } p_i^S = 0 \text{ for } i \notin S$$

If p^0 is optimal, it can be shown that there is a subset S^0 of the index set such that

$$p_i^0 = p_i^{S^0} \text{ for all } i$$

Thus p^0 can be obtained by projecting Z onto the hyperplane spanned by the B_i , $i \in S$ for some S . This method also permits a reduction in the number of projections required to compute the optimum. (For more detail, see the appendices of Refs. 1 and 3.)

Note that the problem of estimating p when the M_i are equal can be expressed entirely in geometrical terms. The vectors A_i can be regarded as points in n -space. The set of all possible mixture signature means $\sum p_i A_i$, where $p = (p_1, \dots, p_m)$ and p is a proportion vector, is the convex hull of the A_i and is called the signature simplex. The covariance matrices M_i can be regarded geometrically as hyperellipsoids centered at the respective vertices A_i of the simplex. Similarly the vectors $B_i = L^{-1} A_i$ are points in n -space and form a geometrical figure called the transformed signature simplex, which is the convex hull of the B_i . The transformed covariance matrices are all-identity and represented by unit spheres centered around the B_i . The point B_p corresponding to the optimal p is the closest point in the transformed simplex to the point $Z = L^{-1} y$. The convex hull of the B_i will be the same as the geometrical figure whose edges connect the B_i in space if the $(n + 1)$ -dimensional vectors $(1, B_i)$ are linearly independent.

3.3. DETECTION OF ALIEN OBJECTS

The maximum likelihood estimate of p obtained by the methods described in previous text may be greatly in error if the area contains alien objects not represented in the signature set. (Objects not represented in the signature set are called alien objects.) To assure fairly accurate estimates of the proportions, it is essential that pixels which include large portions of alien material be eliminated from consideration. Statistical tests have been developed for detecting such points, using the geometric concept of the signature simplex.

It is not possible, without additional information, to estimate the proportions of unspecified materials in a mixture. Nor is it possible to determine unquestionably whether a data point represents a mixture of the specified materials. This can be done only in a statistical sense. If a data point does represent such a mixture, we would expect it to lie fairly close to the signature simplex; just how close will depend on some statistical threshold.

Some alien objects are easier to detect than others. Overall ease of detection depends on (1) the actual proportion of the alien material present and (2) differences in spectra between the alien materials and the materials specified for the mixture. This dependence is true no matter what statistical threshold is used.

We will now briefly discuss the manner in which alien objects are detected. (Appendix I contains more details on the implementation of the alien object tests). The approach taken can be explained in purely geometric terms. The locus of mixture signature means is the signature simplex. Indeed, if there were no statistical variation in the data, then every data point representing a mixture of the specified materials would lie in the simplex. With statistical variation, every such data point will lie "close" to the simplex. This is the key idea.

For the sake of mathematical simplicity, the analysis that follows is based entirely on the transformed simplex and data point Z . However, the alien object test computations carried out as a part of the proportion estimation program employ only the untransformed data vectors.

The alien object test is divided into two parts. The first part, the hyperplane criterion, deals with the projection of Z onto the linear variety determined by the signature simplex. The second part, the out-of-plane test, tests the magnitude of the component of Z lying outside this linear variety. Both are used to determine whether Z is "too far" from the signature simplex.

3.3.1. HYPERPLANE CRITERION

Let y be the data vector, $Z = L^{-1}y$, and let B_i be the vertices of the transformed simplex. Since we are dealing with the projection of Z onto the space spanned by the B_i , we can assume without loss of generality that Z lies within this space. The hyperplane criterion will determine whether a given point Z lies within s "standard deviations" of the locus of mixture signature means. The locus of mixture signature means is the signature simplex. If this is the simplex of transformed means, the locus of points within s standard deviations of any mixture signature mean is approximated by a simplex similar to but larger than the signature simplex. Its faces are parallel to the faces of the signature simplex and a distance s units from them, respectively. This is the "extended signature simplex s units away" which is shown in Fig. 5 for two dimensions and three signatures. The vertices (B_1 , B_2 , and B_3) of the solid-line triangle represent the signature means while the triangle itself describes the locus of mixture signature means. The dashed-line triangle is the extended signature simplex. If the transformed data point Z is outside the extended simplex with $s \geq 2$, there is a low probability that y represents a mixture of only the elements whose means form the signature simplex. Thus the corresponding pixel likely contains some alien materials so proportions should not be estimated for that pixel. The point Z lies within the extended simplex if and only if it is on the same side of each face (hyperplane) as the corresponding opposite vertex of the signature simplex. By using the hyperplanes through the faces, it is easy to determine whether Z is inside or outside. If some preliminary computations are first carried out, this test is easy to implement.

3.3.2. OUT-OF-PLANE TEST

The hyperplane criterion, in actual fact, tests only the projection of Z onto the hyperplane determined by the vertices B_i . But it is still possible, even if Z satisfies that criterion, for Z itself to be far away from the signature simplex, indicating that Z probably represents a mixture containing much alien material. In such case the hyperplane criterion is not an adequate test. Therefore it is desirable to test those components of Z which are orthogonal to the hyperplane determined by the B_i . Such a test has been formulated and is called the out-of-plane test. (See Appendix I for more details.)

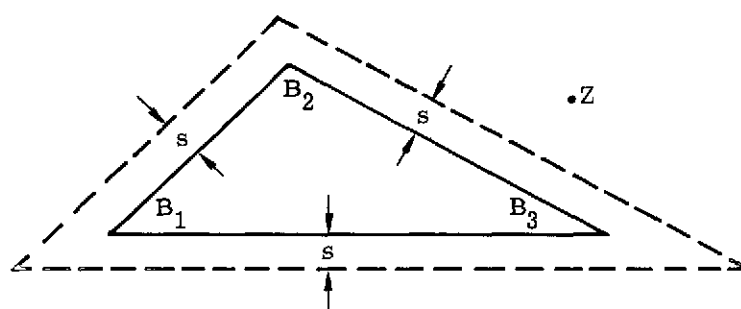


FIGURE 5. THE EXTENDED SIGNATURE SIMPLEX

3.3.3. ALIEN OBJECT TEST SUMMARY

To summarize: The hyperplane criterion determines whether the in-plane projection of Z is more than s standard deviations from the simplex. The out-of-plane criterion tests whether any out-of-plane components of Z are more than $s\sqrt{\phi}$ away from the simplex, where ϕ is the "out-of-plane factor" specified in the control input. The number ϕ is specified by the user and provides a certain flexibility in testing for alien objects.

The alien object test, if s and ϕ are properly chosen, can be very useful in eliminating data points which will corrupt estimates of the proportions. However, like any other statistical test, its validity depends on choice of these threshold values. For the experiment described in this report, s and ϕ were chosen empirically in order to reject a fixed percentage of the training set data points.

The computations required to implement both the hyperplane criterion and the out-of-plane test have been much simplified in the interest of minimizing the amount of computer time per data point. Most of the computation is done once for each rectangular area in a subroutine of the proportion estimation program. The point-by-point computation then consists of two matrix multiplications plus a few other simple operations. On the Control Data C1604 digital computer the processing time required is 0.0167 sec/point.

3.4. DESCRIPTION OF THE PROPORTION ESTIMATION PROGRAM

By using the procedures described in the previous two sections, a program named MIXMAP has been written to estimate the proportions of objects from multispectral data. Inputs consist of (1) control information, (2) the required signatures, and (3) the data from which proportions are to be estimated. Outputs consist of (1) estimated proportions and number of alien objects detected in the area of interest, and (2) a series of digital maps displaying the approximate proportions of each crop in each resolution cell and a map of the alien objects detected.

The MIXMAP program affords the user a number of options, each specified to the computer by the control input. One option is to estimate the proportions "with averaging" over each rectangular area specified to the program. Normally, proportions are estimated separately from each resolution element or data point within an area and then averaged to obtain estimated proportions for the entire area. In application of the averaging option, however, MIXMAP will first average the data vectors and then estimate proportions for the entire area from the averaged data vector. (Reference 3 discusses this technique.)

Another option is the use of "covariance factors." Under this option, scaling of the covariance matrices is made possible, should the user desire to do so. If this option is not used the covariance factors are all set to unity.

MIXMAP is capable of testing for and eliminating alien objects from consideration. This also is a processing option. When the option is exercised, two numbers are required to

specify the threshold (statistical) level of the test: One is the number s of standard deviations of tolerance; the other is an "out-of-plane factor" ϕ . In this option every data point is tested individually, whether the averaging option is used or not.

The control input, in addition to its specification of options, must include: (1) number of pure signatures, (2) number of spectral channels used in defining the signatures, (3) subset of channels to be used in estimating proportions, (4) names of the pure signatures, and (5) the standard data-processing inputs (number of files to skip on the data tape, and the line and point limits for each rectangular area to be processed).

When the averaging option is not used, MIXMAP will provide a series of maps showing the location of the alien objects detected and, for each pixel, the approximate proportions of each material for which a pure signature is provided.

Figure 6 shows examples of the maps generated when carrying out standard (one class for one point) classification and proportion estimations of each class for each point. The standard classification map appears in the center of the group of figures. Here a distinct symbol is assigned to each class; then as each point or resolution element is classified, the corresponding symbol is printed in the appropriate location. For this map the classes corn, soybeans, bare soil, alfalfa, cut alfalfa, and alien are represented by the symbols θ , X, \times , =, -, and blank, respectively. When estimating proportions one map is generated for each class with symbols now assigned to represent the proportion of that class at each point. For purposes of this illustration the proportion ranges of 0-20%, 20-60%, and 60-100% are represented by blank, *, and \star , respectively. Of course, as many ranges from 0-100% may be designated and mapped as there are distinct symbols.

Only the alien object map differs from what has just been described. For this map proportions are not used and all points to be designated as alien are assigned a particular symbol. An alternate method is to assign several distinct symbols, each specifying a range of "distance" of the point in question from the vector space spanned by the signatures.

3.5. SIGNATURE ANALYSIS

In our previous work we developed a technique for determining whether good estimates of proportions could be obtained with a given set of signatures. This is an important consideration for, with some signature sets (as for example, where one material is spectrally similar to a mixture of the others), the estimates will be poor at best. Whether good estimates can be obtained depends on the shape of the signature simplex. To minimize the error in distinguishing between pure objects with means A_i , the distances between the signature means A_i should be large relative to the spreads (covariances). This requirement applies also to the mixtures problem. But to obtain good estimates of proportions still another condition must be satisfied:

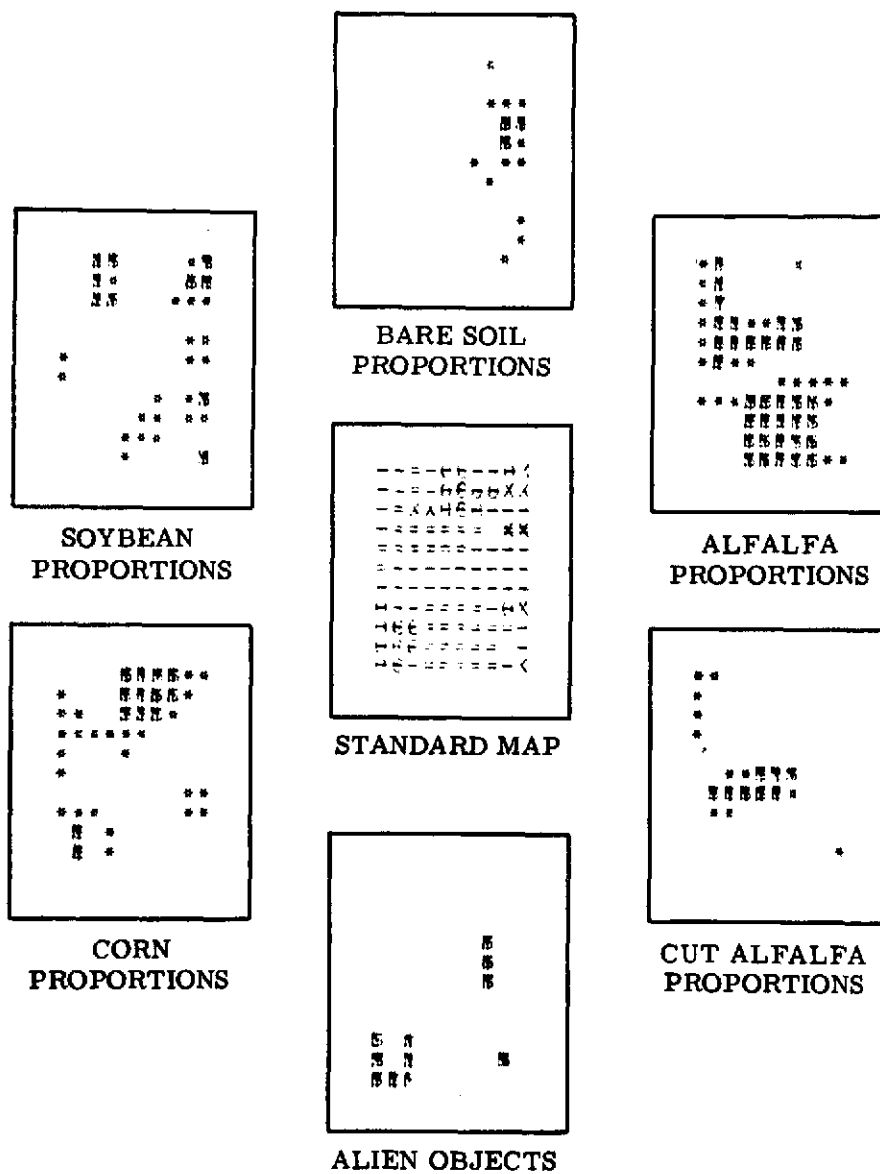


FIGURE 6. EXAMPLES OF STANDARD AND ESTIMATED PROPORTION CLASSIFICATION MAPS

namely, that no vertex of the signature simplex is "too close" to the convex hull of the other vertices. If this last condition is not satisfied, one object might be improperly interpreted as a mixture of other objects. The likelihood of such occurrence can be foreseen by geometric signature analysis.

The key to such analysis is that the signatures and the relationships between them can be viewed entirely in geometrical terms. The signature simplex has already been defined (see Section 3.3.1) as the geometrical figure in signal space whose edges connect pairs of pure signature means. In the nondegenerate case, the one of interest here, each pure signature comprises a distinct vertex of this simplex. The covariance matrices can be interpreted in terms of loci of constant probability. For a Gaussian distribution these loci are hyperellipsoids, centered about the signature mean. One of these ellipsoids, called the unit contour ellipsoid, is chosen for each signature. These contour ellipsoids can be depicted with the signature simplex. A specific geometrical configuration for three signatures and two channels is shown in Fig. 7. In general, if A_i is the signature mean and M_i is the covariance matrix, the unit contour ellipsoid is the set of vectors v satisfying

$$\left\langle v - A_i, M_i^{-1}(v - A_i) \right\rangle = 1$$

Thus the geometrical configuration is completely determined by the A_i and M_i .

Inaccuracies in the estimate of the proportion vector are likely if any vertex of the signature simplex is too close to the opposite face (convex hull of the remaining vertices). What is "too close" will obviously depend on the size and shape of the unit contour ellipsoid about that vertex and hence on the corresponding covariance matrix. Figure 7 depicts a configuration in which the vertices are well separated, in a probability sense, from their opposite faces. Figure 8 on the other hand, illustrates an ill-conditioned configuration in which some points on the face (segment) A_2A_3 are close in a probability sense to the vertex A_1 .

Geometrical signature analysis consists of computing a relative distance r_i from each vertex A_i of the signature simplex to its opposite face. In general, r_i is a measure of how far one signature is from the convex hull of the others, and may be regarded as a probability distance. If all the r_i are large, estimates of the proportion vector should be reliable. On the other hand, if any r_i is small, significant estimation errors may result. The number r_i can be obtained in three steps:

- (1) Find a matrix P such that the i -th contour ellipsoid transformed by P^{-1} is a unit circle about A_i .
- (2) Transform all the A_j by P^{-1} .
- (3) Using the simplex formed by the transformed A_j , let r_i be the minimum distance from A_i to its opposite face.

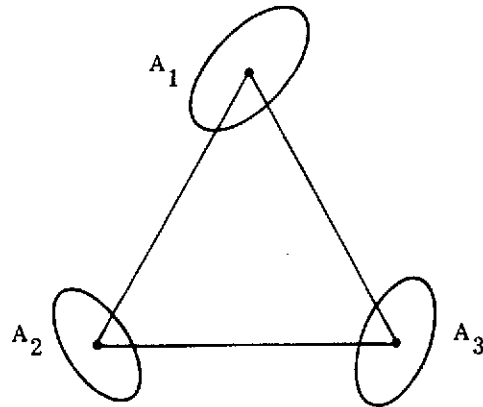


FIGURE 7. SIGNATURE SIMPLEX WITH UNIT CONTOUR ELLIPSOIDS

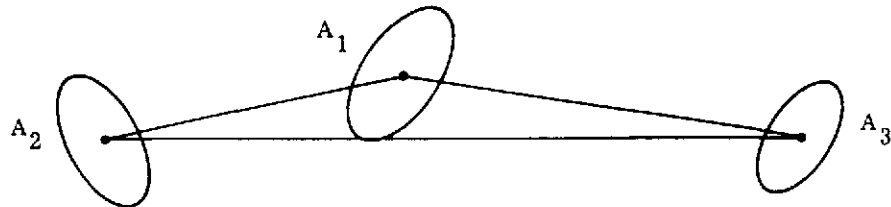


FIGURE 8. ILL-CONDITIONED SIGNATURE SIMPLEX

More specifically, this involves the following computational steps:

- (1) Compute by Cholesky decomposition a matrix such that $PP^T = M_1$.
- (2) Compute the transformed signature mean vectors $B_j = P^{-1}A_j$.
- (3) Let C_i be the closest point to B_i in the convex hull of the other B_j . (This may be computed by the Theil and van de Panne method.)
- (4) Compute $r_i = \|A_i - C_i\|$.

4

APPLICATION OF THE PROPORTION ESTIMATION CAPABILITY

To assess the usefulness of the techniques described in Section 3, a set of multispectral scanner data gathered from aircraft altitudes was selected for testing. (Data gathered from space altitudes would have been preferred, but none was available.) The data set selected was gathered by The University of Michigan multispectral scanner on 21 August 1970 at a flight altitude of 10,000 ft over an agricultural area in Lenawee County, Michigan. Twelve channels of data were recorded with the spectral sensitivities noted in Table 1. Ground truth data had been secured in conjunction with this flight [4].

The remainder of this section describes the preparation and processing of this data and discusses the results achieved in applying the proportion estimation techniques.

4.1. DATA PREPARATION

For proper interpretation of the processing results obtained, the various steps employed in preparing and processing the data should be understood. In this section we detail those necessary steps.

Because the data were initially recorded in analog form, the first steps consisted of deskewing and digitizing the data. Deskewing, an operation that eliminates any channel-to-channel misregistration, is accomplished prior to digitizing by utilizing electronic delay lines to bring the data into register. (Even though the spectrometer channels are aligned optically, the skew problem still exists because of slight differences generally present in the tape recorder record and playback head alignment.)

In digitizing the data we took advantage of certain data collection characteristics to reduce noise effects. From an altitude of 10,000 ft, successive scan lines overlapped a considerable amount. Therefore, rather than digitizing only selected scan lines and sampling those lines at a rate producing a density equivalent to the system optical resolution of 3.3 mr (milliradians), we sampled at an equivalent 5 mr rate using the appropriate electronic response to smooth the data in the scan direction, and averaged several of the scan lines to cover the same ground area which would be covered by a scanner of 5 mr optical resolution.

The next steps in data preparation were to clamp and scale the data. The purpose of both of these operations was to reduce, as much as possible, variations in the data caused by changes in collection and recording system characteristics during the data collection flight. One of these changes might be a change in the recorder offset value. Changes in offset were accounted for by clamping the data to the signal generated during each scan line as the scanner views the dark

TABLE 1. SPECTRAL SENSITIVITY OF THE
MICHIGAN MULTISPECTRAL SCANNER FOR AUGUST 1970

Spectrometer Channel Number	50% Response Points (μm)	Peak Response Points (μm)
1	0.41-0.43	0.42
2	0.43-0.455	0.44
3	0.455-0.47	0.46
4	0.47-0.485	0.475
5	0.485-0.500	0.49
6	0.500-0.520	0.51
7	0.520-0.545	0.535
8	0.545-0.580	0.56
9	0.580-0.63	0.605
10	0.63-0.68	0.65
11	0.68-0.74	0.71
12	0.75-0.855	0.805

interior of the scanner. Changes in system gain may also occur; these are compensated by using, as an automatic gain control, the signal generated when the scanner views the fixed radiance of a reference source.

Upon our completion of these preparatory steps the data were examined to see whether any systematic variations remained. Not too surprisingly, there was a systematic variation in the average scanner signal level as a function of scanner view angle. This effect probably resulted from two factors: the varying effects of the atmosphere as the optical path length from the sensor to the ground varied with changing view angles, and the bidirectional reflectance characteristics of the crops. Since variable effects of this sort would be much reduced in the satellite-based scanners with their smaller total field of view, the systematic angular variations in the Lenawee County data were eliminated using the ACORN IV preprocessing technique developed by ERIM personnel [6-9].

The final step in the data preparation phase consisted of generating the space-simulated data. This was accomplished by averaging data points in both dimensions (across and along the flight path) to produce single pixels or resolution elements measuring 300 ft on a side. This approximates the instantaneous ground patch which will be viewed by the multispectral scanners carried by the ERTS-1 and SKYLAB satellites.

4.2. SIGNATURE EXTRACTION

Before processing the space-simulated data it was necessary to establish the characteristics of the signatures for the major crops and ground covers in the scene so that the computer could be trained to recognize these objects.

From the ground truth information it was determined that the major constituents in the scene were corn, soybeans, bare soil, grain stubble, alfalfa and cut alfalfa. These comprised more than 90% of the scene. In order to locate in the data specific areas containing samples of these items, a graymap of one spectral channel was generated. (A graymap is a pictorial representation of the scene in which specific ranges of scanner signals are assigned distinct printer symbols.) From the graymap it was clear that there would be difficulties in locating a sufficient number of pure resolution elements to provide samples of each crop type for the establishment of crop signatures. (The number of such samples required is discussed in Appendix I; a possible means for extracting pure signatures from data with no pure samples is described in Section 5.) The difficulties mentioned above were a result of the relatively low spatial resolution in the space-simulated data compared with the dimensions of most fields in the scene. This was the case even though Lenawee County has larger fields than most other counties in Michigan.

A graymap of one spectral channel of the 10,000 ft data was generated and several fields of each of the major ground covers were located. Data points were extracted from each of these fields and signatures were calculated for each field. The signatures thus obtained were evaluated via geometric signature analysis (see Section 3.5); based on this analysis, those considered to be unrepresentative samples were eliminated from further consideration. The remaining signatures in each crop group were then combined to obtain a final crop signature. The final signature set consisted of signatures for corn, soybeans, bare soil, alfalfa, and cut alfalfa.

4.3. DATA PROCESSING AND EVALUATION

Before actually processing the set of space-simulated data, we applied the proportion estimation technique to the training fields in the 10 Kft data in order to estimate the accuracy that may be expected in identifying the pure elements of the "space data." The results, as given in Table 2, show that 80-90% of the data points were correctly classified, with the remainder classified either as one of the other four materials or as an alien object. These results are similar to those obtained using the conventional classification algorithms on other data sets. Therefore, whether we use conventional recognition or the proportion estimation algorithm, we can expect comparable accuracy for those data points or pixels containing only one object class.

We shall now discuss the results of estimating proportions from the space-simulated data. The entire data set was processed using three methods: (1) the conventional classification approach wherein each pixel is classified either as being an alien object or one, and only one, of the five classes, (2) the proportion estimation approach in which each pixel is classified as being an alien object or a combination of the five classes, and (3) the proportion estimation approach with averaging where (a) each pixel is classified as to whether or not it represents an alien object and (b) all non-alien object points are averaged and the proportions then estimated from the single averaged data point.

TABLE 2. RESULTS OF ESTIMATING PROPORTIONS
ON TRAINING AREAS OF 10K FT, AUGUST
1970 — LENAWEE COUNTY DATA

Material	Correctly Classified (%)	Incorrectly Classified (%)	Classified as Alien (%)
Corn	88.0	12.0	2.5
Soybeans	84.8	15.2	1.9
Bare Soil	90.4	9.6	4.6
Alfalfa	80.3	18.7	1.7
Cut Alfalfa	82.0	18.0	3.4

Table 3 summarizes the results of classifying the entire space-simulated data set. Although ground truth information was not available for the entire scanned scene, we still find the results (Table 3) quite interesting. At first glance the results of the three methods look so similar that use of any approach more complicated than standard classification algorithms might well seem questionable. Certainly the results for corn are essentially the same; but as one scans down the table the differences become as great as 14% for cut alfalfa. However even this large a difference might not be too important if the truth lay between the two estimates.

We decided to examine in more detail the results for certain areas of the scene where reliable ground truth was available. Some fairly typical findings of this examination are shown in Figs. 9 and 10; these findings are discussed and some general conclusions drawn in the following paragraphs.

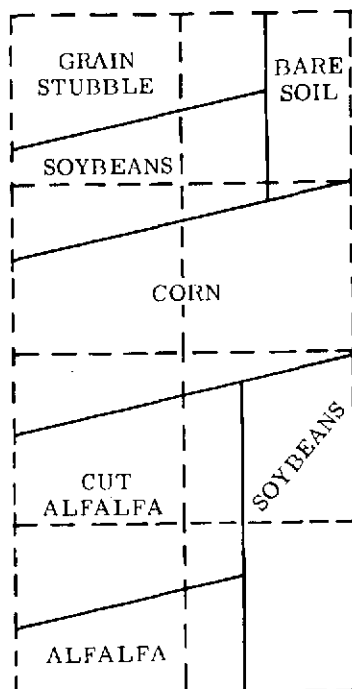
Figure 9 presents the results for an area containing each of the five major classes as well as an alien object—namely, a field of grain stubble. As in Fig. 1 the dashed-line squares represent individual pixels. In this example all pixels included at least two object classes. Note that the field boundaries are skewed with respect to the pixels. This is because the scanner aircraft flew at a crab angle of several degrees to correct the flight path for effects of a crosswind.

On examining the table included as a part of Fig. 9 it is obvious that the results achieved in estimating proportions are much more accurate than those using the standard classification algorithm. Even though 26% of the area actually contained soybeans the standard algorithm recognized no soybeans; at the same time, the amounts of corn and cut alfalfa were grossly overestimated. When estimating proportions on the average point, the cut alfalfa was also overestimated. No alien elements were identified by any of the methods—possibly because the alien object, a grain stubble field, had low weeds and a radiance spectrum resembling that of cut alfalfa. Overall, the point-by-point proportion estimates produced the most accurate results.

The area presented in Fig. 10 also includes all five major classes as well as some alien objects: the farmstead, the road, and the idle field. Some of the fields were large enough to fully contain one or more pixels within their boundaries. The accuracy of the standard classification technique was once again quite poor. Soybeans and bare soil were underestimated whereas the cut alfalfa was greatly overestimated. A total of nine pixels straddled the boundaries between fields of bare soil and soybeans. Upon examining these pixels we found that each was classified as cut alfalfa by the standard classification technique. But when we used the proportion estimation algorithm, the alien objects were overestimated. This is understandable since each point was classified as being either totally alien or a combination of the other five classes; and if more than two of the elements containing the road were classified

TABLE 3. PROPORTIONS OF MATERIALS FOR AUGUST 1970
LENAWEE COUNTY DATA SET USING THREE CLASSIFICATION
APPROACHES AFTER COMBINING DATA POINTS TO SIMULATE
300 × 300 FT RESOLUTION

Material	Conventional Point by Point Classification (one class per point)	Point by Point Proportion Estimates (as many as five classes per point)	Proportion Estimates on a Single Point Representing the Average of Non-alien Points in the Scene
Corn	0.264	0.248	0.253
Soybeans	0.129	0.196	0.158
Bare Soil	0.034	0.101	0.082
Alfalfa	0.269	0.173	0.207
Cut Alfalfa	0.249	0.109	0.127
Alien	0.055	0.173	0.173



MATERIAL	GROUND TRUTH	STANDARD CLASSIFICATION (PT BY PT)	PROPORTION ESTIMATES (PT BY PT)	PROPORTION ESTIMATE ON AVG PT
CORN	0.262	0.500	0.380	0.347
SOYBEANS	0.262	0.000	0.199	0.220
BARE SOIL	0.069	0.000	0.074	0.019
ALFALFA	0.094	0.125	0.091	0.029
CUT ALFALFA	0.194	0.375	0.256	0.386
ALIEN	0.119	0.000	0.000	0.000

FIGURE 9. COMPOSITION OF A PORTION (AREA 1) OF SPACE-SIMULATED LENAWEE COUNTY DATA AS DETERMINED BY FOUR METHODS

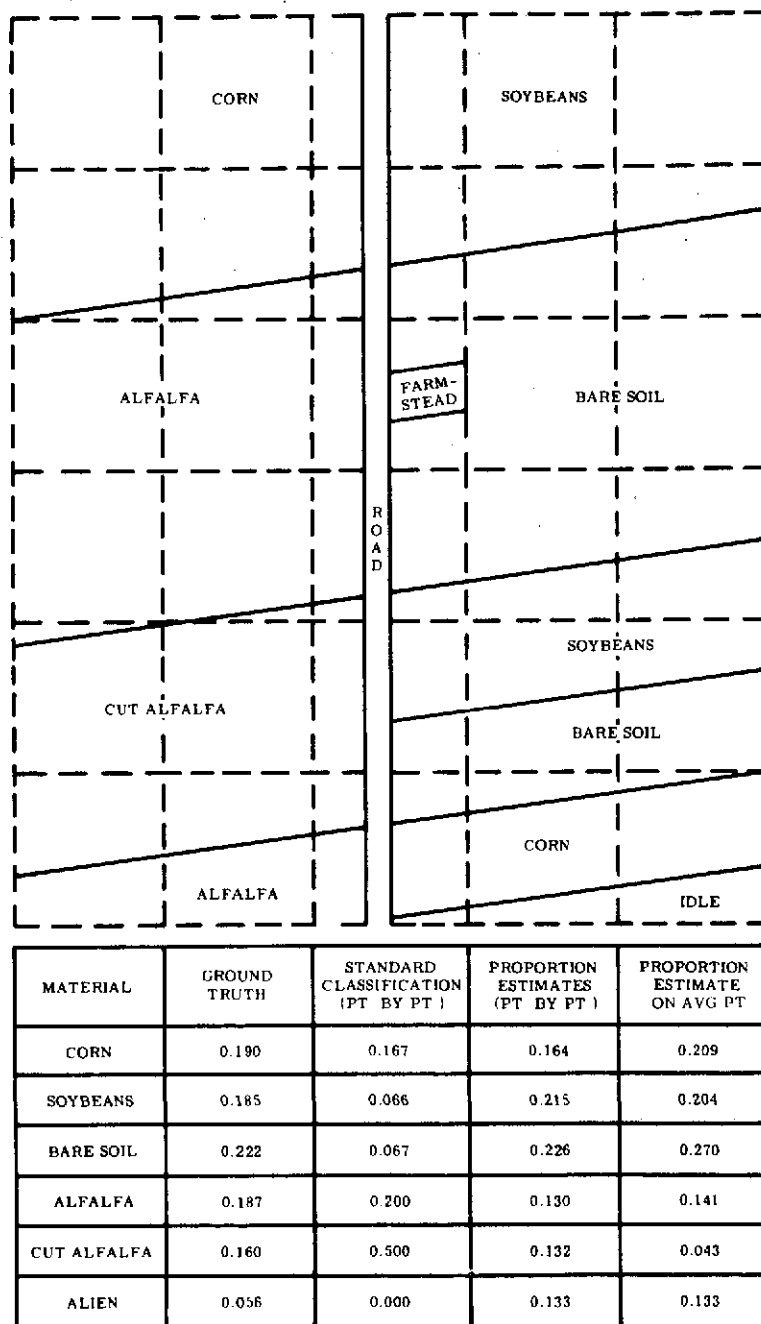


FIGURE 10. COMPOSITION OF A PORTION (AREA 2) OF SPACE-SIMULATED LENAWEE COUNTY DATA AS DETERMINED BY FOUR METHODS

as alien, that category would be overestimated. The best overall results were again achieved by estimating proportions on a point-by-point basis; estimating proportions from the average point was a close second.

Several other areas were examined with results similar to those already described. Two observations seemed to hold true for almost all the areas: the mean square error of the point-by-point proportion estimates was much less than for the standard classification; and the errors for the standard classification were both positive and negative. Perhaps these traits explain how the proportions of materials estimated for the entire scene and listed in Table 3 could be so similar. Apparently the scene was so structured that the standard technique erred almost equally on both sides of the truth.

In sum, our results indicate that the proportion estimation technique provided a better estimate of crop area in those Lenawee County test regions for which reliable ground truth was available—somewhat better than the conventional classification algorithm. It would seem that proportion estimation would be the better technique for any agricultural area.

5

EXTRACTING SIGNATURES FROM LOW RESOLUTION DATA

Our experience with the single set of space-simulated data showed that, when operating on data gathered at space altitude, some difficulty may be encountered in obtaining signatures in the usual manner since there may not be a sufficient number of pixels containing pure samples of the materials to be recognized (see Appendix I). Indeed it is possible that for some regions, every pixel will contain a mixture. Hence a method is needed for obtaining signatures from data representing mixtures of the specified materials.

It is desired to compute the signatures from pairs of data vectors y and proportion vectors p (obtained from ground truth). A general solution to this problem has not been obtained, but we found a mathematical solution for the case where the covariance matrices are equal. This solution is described in the following paragraphs.

Let m be the number of materials or signatures to be obtained, n the number of spectral channels, and N the number of samples (pixels) for which multispectral data and proportion vectors are available. Assume that $N \geq m$ and that there is a common covariance matrix M for the signatures to be estimated. Assume also that the N samples are statistically independent. Then the m by n matrix A , whose rows are the signature mean vectors, may be estimated by least squares. Let

$$Q = \sum_{k=1}^N ||p^k A - y^k||^2$$

where y^k is a data vector and p^k is the corresponding ground truth proportion vector. Also let

$$P = \begin{bmatrix} \text{---} p^1 \text{---} \\ \vdots \\ \text{---} p^N \text{---} \end{bmatrix}, Y = \begin{bmatrix} \text{---} y^1 \text{---} \\ \vdots \\ \text{---} y^N \text{---} \end{bmatrix}$$

We may write

$$Q = \sum_{i=1}^n Q_i, \text{ where } Q_i = \sum_{k=1}^N \left(\sum_{j=1}^m p_j^k A_{ji} - y_i^k \right)^2$$

and minimize each Q_i separately. This leads to n least-squares problems, each to estimate m quantities A_{ji} . It follows that

$$\frac{\partial Q_i}{\partial A_{ji}} = 2 \sum_{k=1}^N p_j^k \left(\sum_{\ell=1}^m p_\ell^k A_{\ell i} - y_i^k \right) = 2[(P^T P)A - P^T Y]_{ij}$$

Provided that all $A_{ji} > 0$, a necessary condition for the optimal estimate is

$$\frac{\partial Q_i}{\partial A_{ji}} = 0 \text{ for } i = 1, \dots, n; \quad j = 1, \dots, m$$

so that $(P^T P)A = P^T Y$. Thus the equation for estimating A is

$$A = (P^T P)^{-1} P^T Y$$

When A has been estimated, it is also possible to estimate the covariance matrix M . Since the pure signature covariances are equal, all mixture signature covariances are independent of p^k . Let \bar{X} be a matrix whose rows X_i are data vectors from the m pure signature distributions. Then

$$\text{cov}(p^k \bar{X}) = \sum_{i=1}^m p_i^k \text{cov } X_i$$

$$\sum_{i=1}^m p_i^k M = M \left(\sum_{i=1}^m p_i^k \right) = M$$

Thus M can be estimated by the sample covariance

$$S = \frac{1}{N-1} \sum_{k=1}^N (p^k A - y^k)(p^k A - y^k)^T$$

The above approach has not been tested but certain difficulties are obvious: suitable estimates of A and S can be obtained only if (1) enough data (p^k, y^k) are available and (2) the ground truth values p^k are sufficiently accurate. This stringency may necessitate changes in data handling as well as in the procedure for obtaining ground truth. Even for well known areas, it will be difficult to locate the boundaries of fields relative to boundaries of individual pixels in data gathered from space altitudes.

6

CONCLUSIONS AND RECOMMENDATIONS

As discussed earlier in this report, there is a need for multispectral data processing methods whereby the proportions of objects and materials appearing within the instantaneous field of view of a scanning system may be estimated. Toward filling this need, we have developed some promising techniques and tested them to determine their utility in estimating crop acreage from space.

The test results were quite encouraging in that our proportion estimation techniques produced more accurate crop acreage estimates than did a parallel application of the conventional classification technique. Certain problems still exist, however, in applying these new techniques operationally.

One of these problems is the computer time required to estimate proportions on a point-by-point basis. For an ERTS frame, this time requirement is significantly more than for the conventional classification approach. In its present implementation on the Control Data 1604, more than 100 hours will be required for estimating proportions; this may be compared to 17 hours for the conventional approach. The present implementation employs the method of Theil and van de Panne. Another method we examined (the Complementary Pivot Method) is faster and would be applicable if a problem of numerical instability could be eliminated.

Some work has also been done toward finding an unconstrained solution to the mixtures problem, but to date this approach has produced much less accurate results than we had hoped for.

There is also the problem of obtaining signatures from mixtures data; and, as indicated in Section 5, it may be necessary to derive such signatures at space altitudes. A mathematical solution to the problem has been presented in Sec. 5, however, this solution may not be practical unless certain changes are made in the data handling procedures. The question of optimum thresholds in application of the alien object detection criteria remains unanswered at this time.

In this experiment the alien object threshold numbers s and ϕ were determined empirically in order to reject about 10% of the training set data points. It would have been better, however, to have a definite algorithm for choosing values of s and ϕ ; we feel that this merits further work.

Although the problem areas identified above need additional work, we believe that the test results reported herein are sufficiently encouraging to justify further tests on satellite multispectral scanner data. We plan to carry out such tests on ERTS-1 MSS data in the near future.



Appendix I

NUMBER OF SAMPLES REQUIRED FOR TRAINING

In training a computer to automatically and accurately classify a scene, the signatures used must likewise be accurate. And to ensure accurate signatures (means, variance, and covariances) some minimum number of samples is required. This appendix addresses the number-of-samples problem; it is based on an internal memorandum written by R. Crane and W. Richardson [5].

The derivation of Ref. 5 is based on signal to noise ratio, R . The sample (estimated) covariance is

$$S = (1/N - 1) \sum_{k=1}^N (X^k - \bar{X})(X^k - \bar{X})^T$$

where $\bar{X} = E(X^k)$ and N is the number of samples and the variance of S_{ij} is

$$\frac{\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}}{N - 1}$$

The signal to noise ratio for S_{ij} is approximately

$$R_{ij} = \frac{\sqrt{\sigma_{ii}\sigma_{jj}}}{\sqrt{\frac{\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}}{N - 1}}} = \sqrt{\frac{N - 1}{\frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} + 1}}$$

Since

$$0 \leq \frac{\sigma_{ij}^2}{\sigma_{ii}\sigma_{jj}} \leq 1$$

then

$$\sqrt{\frac{N - 1}{2}} \leq R_{ij} \leq \sqrt{N - 1}$$

The signal to noise ratio for S_{ii} is

$$R_{ii} = \frac{\sigma_{ii}}{\sqrt{\frac{2\sigma_{ii}^2}{N - 1}}} = \sqrt{\frac{N - 1}{2}}$$

The percentage errors (E), representing inaccuracy in the covariance and variance terms, are $100/R_{ij}$ and $100/R_{ii}$ respectively. The upper limit for both of these terms for a fixed N is $100\sqrt{2/N - 1}$. This quantity is plotted in Fig. I-1 for N ranging from 3 to 1000. It is clear that as the number of samples is reduced, the percent inaccuracy increases dramatically. Even for what may seem to be a reasonably large number of samples, the inaccuracy is fairly sizable; for example, 100 samples still allow an inaccuracy of 14%. Obviously, a large number of training samples must be available if the characteristics of the signatures are to be established with desired accuracy.

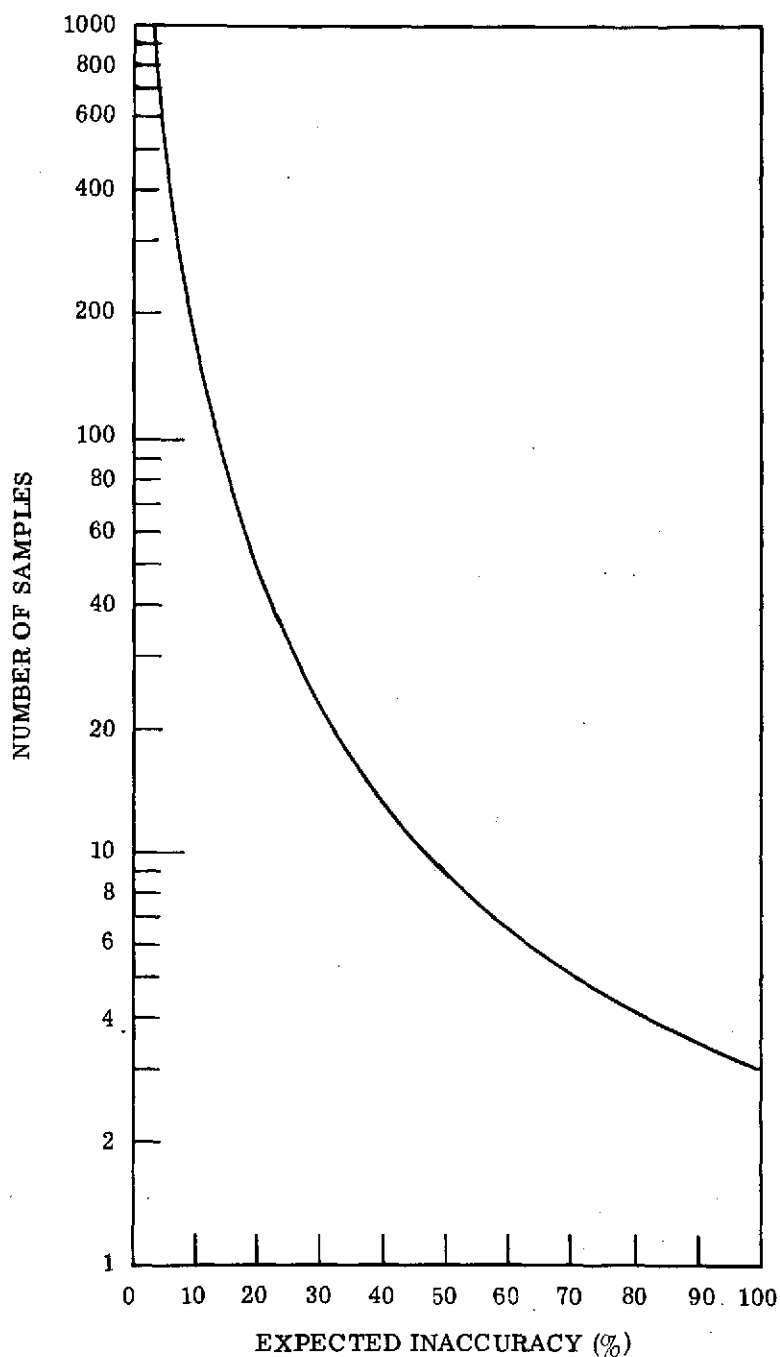


FIGURE I-1: INACCURACY IN THE ESTIMATES OF SIGNATURES AS A FUNCTION OF THE NUMBER OF SAMPLES

Appendix II

DETAILS OF THE ALIEN OBJECT TESTS

The alien object tests were briefly described in Section 3.3. In this appendix we provide more details.

Hyperplane Criterion

Preliminary Computations

First, some coordinate transformations must be carried out. In order to deal with the extended simplex in geometric terms, it is necessary to reduce the dimension of the signature mean vectors from n (number of channels) to $m-1$ (number of signatures or vertices less one). If the signature simplex is nondegenerate, which we assume, then there is an $(m-1)$ -dimensional basis whose space includes all the points of the simplex. This basis is a set of n vectors D_1, \dots, D_{m-1} . These are obtained as follows

Let B_1, \dots, B_m be the transformed signature means.

Define $C_{i-1} = B_i - B_c$ for $i = 2, \dots, m$ where $B_c = (1/m) \sum_i B_i$ is the centroid of the simplex.

Construct the orthonormal basis D_1, \dots, D_{m-1} . The D_1 are defined by

$$D_1 = C_1 / \|C_1\|$$

$$D_2 = [-(D_1 \cdot C_2)D_1 + C_2] / \|[-(D_1 \cdot C_2)D_1 + C_2]\|$$

$$D_3 = [-(D_1 \cdot C_3)D_1 - (D_2 \cdot C_3)D_2 + C_3] / \|[-(D_1 \cdot C_3)D_1 - (D_2 \cdot C_3)D_2 + C_3]\|$$

.

.

.

$$D_k = \left[- \sum_{i=1}^{k-1} (D_i \cdot C_k) D_i + C_k \right] / \left\| \left[- \sum_{i=1}^{k-1} (D_i \cdot C_k) D_i + C_k \right] \right\|$$

.

.

.

Since the D_i are a set of mutually orthogonal unit vectors, the B_i can be represented in the new basis by the $(m-1)$ -vectors V_i defined by

$$V_i = (B_i \cdot D_1, B_i \cdot D_2, \dots, B_i \cdot D_{m-1})$$

Second, certain constants must be obtained for testing each data point y . These, of course, are related to the equations of the hyperplanes that form the faces of the extended simplex. Having worked out these equations, we can formulate a computationally efficient scheme for determining whether a point is inside or outside the extended simplex.

Assuming nondegeneracy of the original signature simplex, none of its face hyperplanes will pass through the origin.* Thus the equation of each k -th hyperplane is of the form $\bar{a}_k \cdot x = 1$, when \bar{a}_k is a normal to the hyperplane and x is any point in it. The face hyperplane opposite the vertex V_k (or B_k) must pass through the vertices $V_1, \dots, V_{k-1}, V_{k+1}, \dots, V_m$. Except for an arbitrary scalar factor we have

$$\bar{a}_k = \begin{bmatrix} \bar{a}_{k1} \\ \vdots \\ \bar{a}_{k,m-1} \end{bmatrix} = \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1,m-1} \\ \vdots & \vdots & & \vdots \\ V_{k-1,1} & V_{k-1,2} & \cdots & V_{k-1,m-1} \\ V_{k+1,1} & V_{k+1,2} & \cdots & V_{k+1,m-1} \\ \vdots & \vdots & & \vdots \\ V_{m1} & V_{m2} & \cdots & V_{n,m-1} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

and the unit normal N_k to the k -th hyperplane is given by $N_k = \bar{a}_k / \|\bar{a}_k\|$. The hyperplane equation is now

$$N_k \cdot X = \frac{1}{\|\bar{a}_k\|}$$

Now consider the equation of the k -th face hyperplane in the extended simplex s units away. Clearly the hyperplane $N_k \cdot X = 1/\|\bar{a}_k\|$ divides the space into two regions

$$R_1 = \left\{ X: N_k \cdot X < \frac{1}{\|\bar{a}_k\|} \right\}$$

$$R_2 = \left\{ X: N_k \cdot X > \frac{1}{\|\bar{a}_k\|} \right\}$$

If $V_k \in R_1$, then the equation of the k -th face hyperplane of the extended simplex is

$$N_k \cdot X = \frac{1}{\|\bar{a}_k\|} + s \equiv h_k$$

Otherwise it is

$$N_k \cdot X = \frac{1}{\|\bar{a}_k\|} - s \equiv h_k$$

*Note that by construction, the origin of the simplex is at its centroid.

The constants N_k and h_k will be used to determine whether the mixture corresponding to a point y is likely to contain a significant proportion of unknown material.

Hyperplane Criterion Applied to Each Point

As noted earlier, a data point or resolution element is to be flagged whenever the point Z lies outside the extended simplex. This criterion is easy to define mathematically. A point Z' is inside the extended simplex whenever it is on the same side of the k -th hyperplane (extended simplex) as V_k , for all k . Further, Z' is on the same side of the k -th hyperplane as V_k if and only if

$$\text{sgn}(N_k \cdot Z' - h_k) = \text{sgn}(N_k \cdot V_k - h_k)$$

The value of $\text{sgn}(N_k \cdot V_k - h_k)$ is a constant independent of data and can be computed in advance of processing. In terms of the observed data point y ,

$$Z' = DL^{-1}y \quad (Z', y \text{ column vectors})$$

where

$$D = \begin{bmatrix} D_1 \\ \vdots \\ D_{m-1} \end{bmatrix}$$

$$L^T L = \bar{M}$$

and where \bar{M} is the average covariance matrix of the signatures. If N_k is a row vector, then

$$N_k \cdot Z' = N_k Z' = (N_k DL^{-1})y$$

The row vector (matrix product) $N_k DL^{-1}$ can be obtained in advance, and need not be computed for every y . It is not necessary to transform y in order to apply the hyperplane criterion.

Programming Details

A subroutine in MIXMAP does all preliminary computations and computes the matrix products $N_k DL^{-1}$ and the constants h_k . Required inputs are s , m , n (number of channels), and the B_i . One additional output is

$$\text{ISIGN} = \sum_{k=1}^m I(k)2^{k-1}$$

where $I(k) = 1$ if $N_k \cdot V_k - h_k < 0$, and 0 otherwise. To check a data point y , check the sign bit of $(N_k DL^{-1} \cdot y - h_k)$ against the k -th bit of ISIGN, for every k . If these bits are different for any k , the point y is flagged. It is also flagged if $\|y - A_c\|$ is large.

Just as the hyperplane criterion deals with the projection of Z onto the space spanned by the D_j , the out-of-plane test deals with the projection of Z onto the orthogonal complement of this space. However, the latter test is simpler in that it is concerned only with the magnitudes of the components of Z orthogonal to the space spanned by the D_j . It is sufficient to complete the basis for n -space, where n is the number of spectral channels.

The dimension of the space spanned by the B_i is $m-1$, where m is the number of signatures. The D_j form an orthonormal basis for this space. Therefore, the D_j form part of an orthonormal basis for n -space. The additional vectors A_1, \dots, A_{n-m+1} form an orthonormal basis for the orthogonal complement of the space spanned by the D_j . The additional vectors can be obtained by a Gram-Schmidt type of process. Let

$$A_K = e_{K+r} + \sum_{j=1}^{m-1} \alpha_j D_j + \sum_{j=1}^{K-1} \beta_j A_j$$

where e_ℓ is a unit vector with ℓ -th component 1 and the others 0, and α_j and β_j are unknown coefficients to be obtained. We want the A_i to have the property that

$$A_i \cdot A_j = 0 \quad \text{if } i \neq j$$

$$A_i \cdot A_j = 1 \quad \text{if } i = j$$

$$A_j \cdot D_K = 0 \quad \text{for all } j = 1, \dots, n-m+1$$

$$K = 1, \dots, m-1$$

i.e., the D_j and A_j together form an orthonormal basis for n -space. The number r is initially 0 but may be incremented by 1 whenever e_{K+r} turns out to be a linear combination of $D_1, \dots, D_{m-1}, A_1, \dots, A_{K-1}$. The desired basis is obtained if

$$\alpha_j = -D_{j,K+r}$$

$$\beta_j = -A_{j,K+r}$$

that is, if the A_K are defined by

$$A_K = e_{K+r} - \sum_{j=1}^{m-1} D_{j,K+r} D_j - \sum_{j=1}^{K-1} A_{j,K+r} A_j$$

for $K = 1, \dots, n-m+1$. If

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_{n-m+1} \end{bmatrix}$$

and Z is a column vector, then AZ is the desired out-of-plane projection.

The hyperplane criterion determines whether the in-plane projection of Z is more than s standard deviations from the simplex. The out-of-plane criterion tests whether any out-of-plane components of z are more than $s\sqrt{\phi}$ away from the simplex, where ϕ is the "out-of-plane factor" specified in the control input. The value of ϕ is specified by the user and thus provides some flexibility in screening for alien objects.

REFERENCES

1. Horwitz, H. M., R. F. Nalepka, P. D. Hyde, and J. P. Morgenstern. 1971. "Estimating the Proportions of Objects Within a Single Resolution Element of a Multispectral Scanner," Seventh International Symposium on Remote Sensing of Environment, May 1971, at Ann Arbor, Michigan.
2. Berge, C., and A. Ghouila-Houri. 1965. Programming, Games, and Transportation Networks, John Wiley.
3. Nalepka, R. F., H. M. Horwitz, and P. D. Hyde. 1972. "Estimating Proportions of Objects From Multispectral Data," Willow Run Laboratories Technical Report 31650-73-T.
4. Nalepka, R. F., H. M. Horwitz, and N. S. Thomson. 1971. "Investigations of Multispectral Sensing of Crops," Willow Run Laboratories Technical Report 31650-30-T.
5. Crane, R. B., and W. Richardson. 1972. "Size of Sample Required to Estimate Mean and Covariance Matrix of a Multivariate Normal Distribution," Willow Run Laboratories Memorandum of 23 March 1972.
6. Kriegler, F. J., W. A. Malila, R. F. Nalepka, and W. Richardson. 1969. "Preprocessing Transformations and Their Effects on Multispectral Recognition," Proceedings of Sixth International Symposium on Remote Sensing of Environment, Oct. 1969, at Willow Run Laboratories of the Institute of Science and Technology, The University of Michigan, Ann Arbor.
7. Crane, R. B. 1971. "Preprocessing Techniques to Reduce Atmospheric and Sensor Variability in Multispectral Scanner Data," Proceedings of Seventh International Symposium on Remote Sensing of the Environment, May 1971, at Willow Run Laboratories of the Institute of Science and Technology, The University of Michigan, Ann Arbor.
8. Nalepka, R. F., J. P. Morgenstern, and W. L. Brown. 1972. "Detailed Interpretation and Analysis of Selected Corn Blight Watch Data Sets," Proceedings of the Fourth Annual Earth Resources Program Review, Jan. 1972, at Manned Spacecraft Center, Houston, Texas.
9. Nalepka, R. F., and J. P. Morgenstern. 1972. "Signature Extension Techniques Applied to Multispectral Scanner Data," Proceedings of Eighth International Symposium on Remote Sensing of Environment, Oct. 1972, at Willow Run Laboratories of the Institute of Science and Technology, The University of Michigan, Ann Arbor.